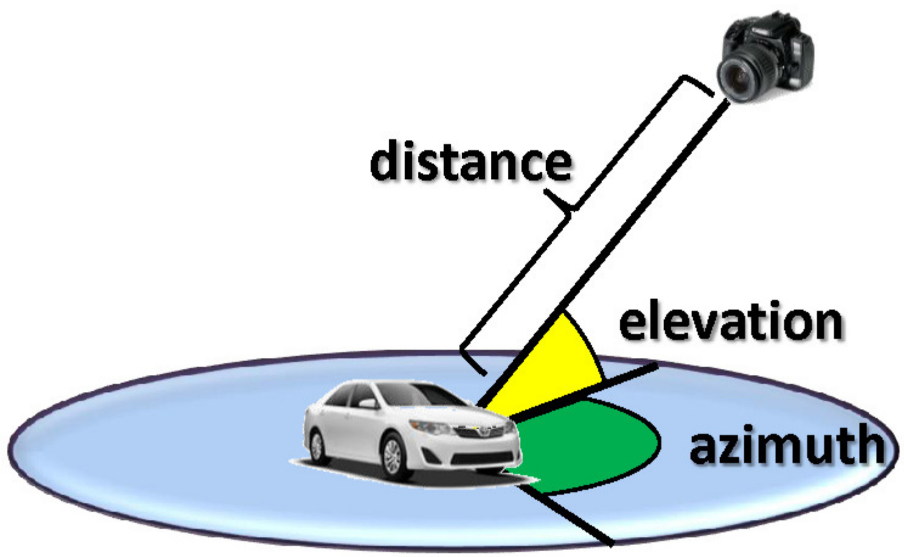
Is 2D Information Enough For Viewpoint Estimation?

Amir Ghodrati, Marco Pedersoli, Tinne Tuytelaars

BMVC 2014

Problem Definition

- Viewpoint estimation: Given an image, predicting viewpoint for object of interest



[1] http://cvgl.stanford.edu/projects/pascal3d.html

Problem Definition

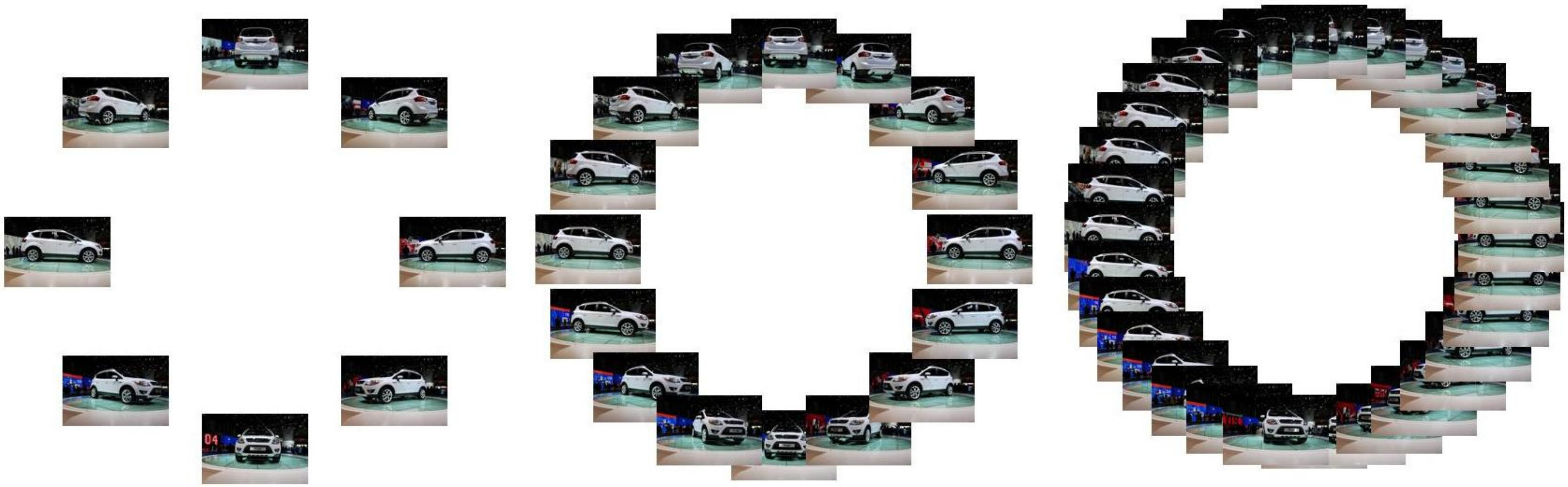
- Viewpoint estimation: Given an image, predicting viewpoint for object of interest

Problem Definition

- Viewpoint estimation: Given an image, predicting viewpoint for object of interest

Problem Definition

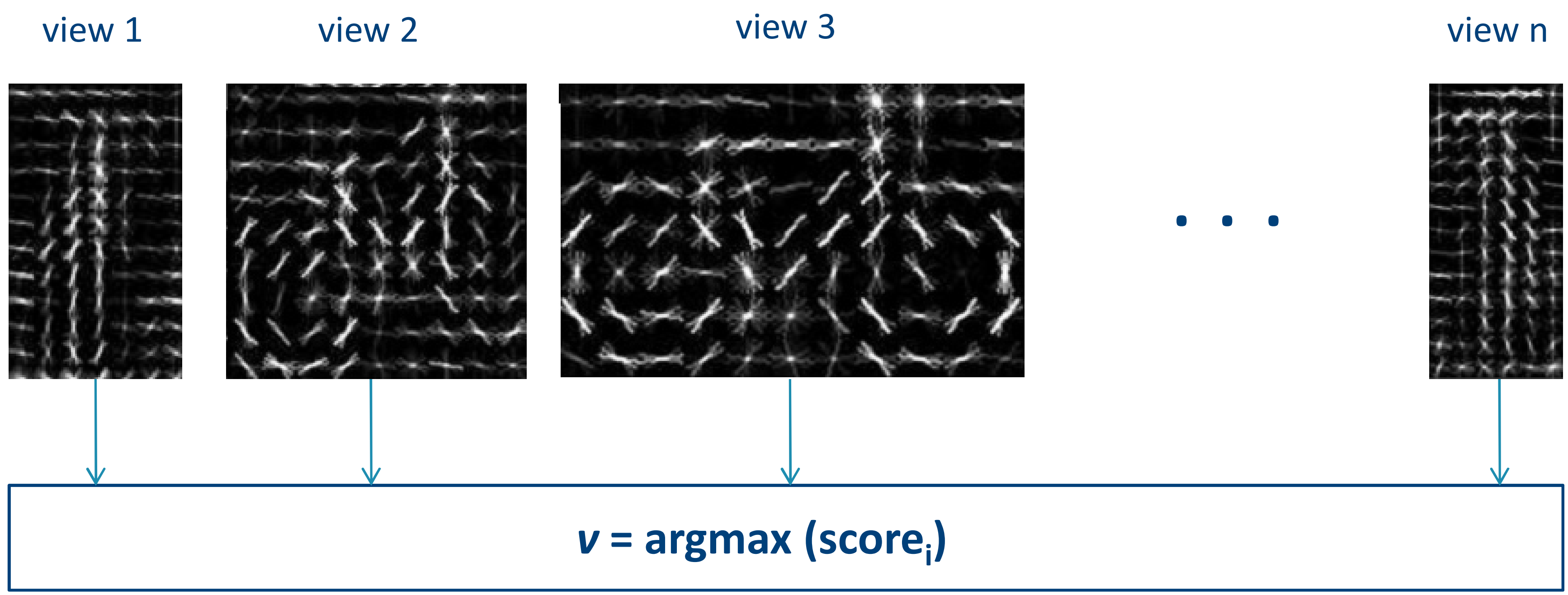
- Viewpoint estimation: Given an image, predicting viewpoint for object of interest



- Fine-grained task of viewpoint estimation

Related works : Detector-based 2D models

- Inspired by detectors that have proven to perform well for the single view case



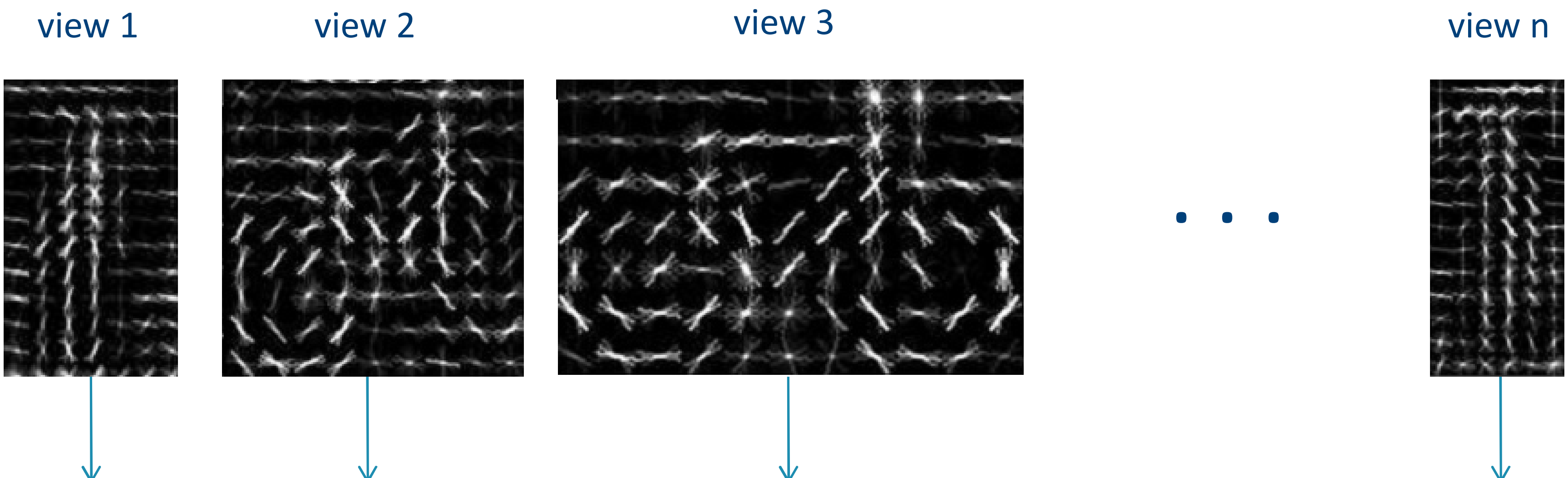
Ch. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In ECCV, 2010.
R.J. Lopez-Sastre, T. Tuytelaars, S. Savarese,: Dpm revisited: A performance evaluation for object category pose estimation. In: ICCV-WS CORP. (2011)

PSI VISICS

KU LEUVEN

Related works: Detector-based 2D models

- Inspired by existing detectors that have proven to perform well



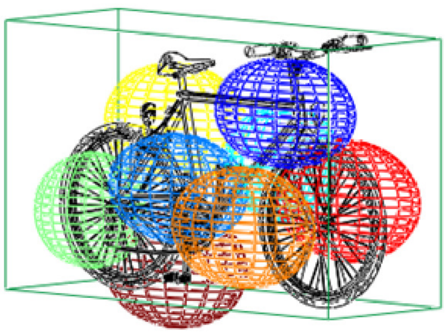
require evaluating a large number of view-based detectors at test time

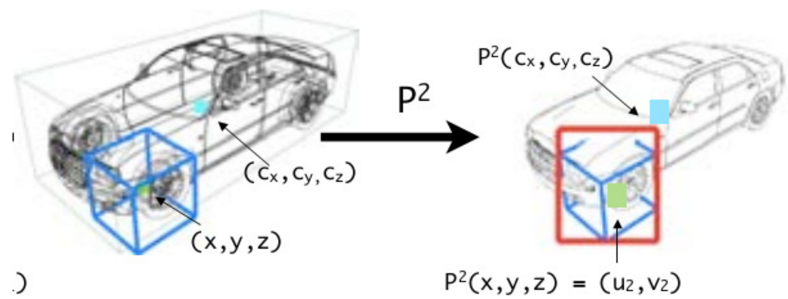
PSI VISICS

KU LEUVEN

Related works: Embrace 3D

- Establish connections between views of an object by mapping them to 3D model.
- 3D geometry is provided in the form of
  - 3D CAD models / Point clouds / Depth sensor
- Performs fine-grained viewpoint estimation





Left: B. Pepik, P. Gehler, M. Stark, B. Schiele. 3d2pm–3d deformable part models. In ECCV, 2012.

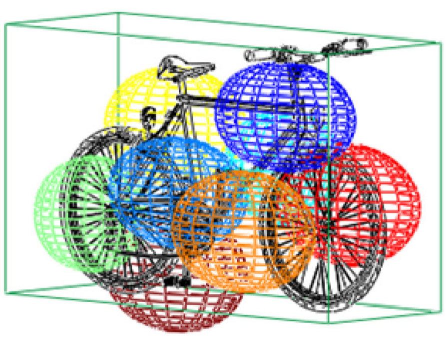
Right: B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In CVPR, 2012

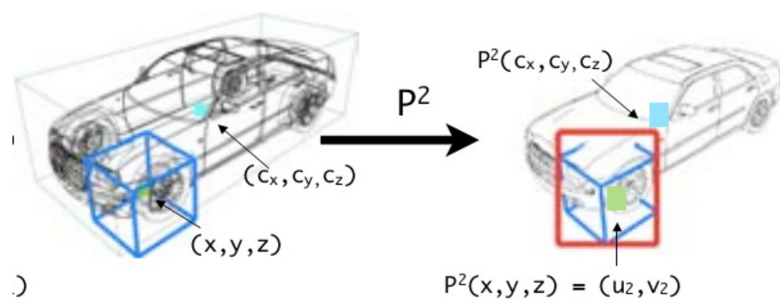
PSI VISICS

KU LEUVEN

Related works: Embrace 3D

- Establish connections between views of an object by mapping them to 3D model.
- 3D geometry is provided in the form of
  - 3D CAD models / Point clouds / Depth sensor
- Performs fine-grained viewpoint estimation





3D information are not always available, for all classes. sometimes are expensive to collect

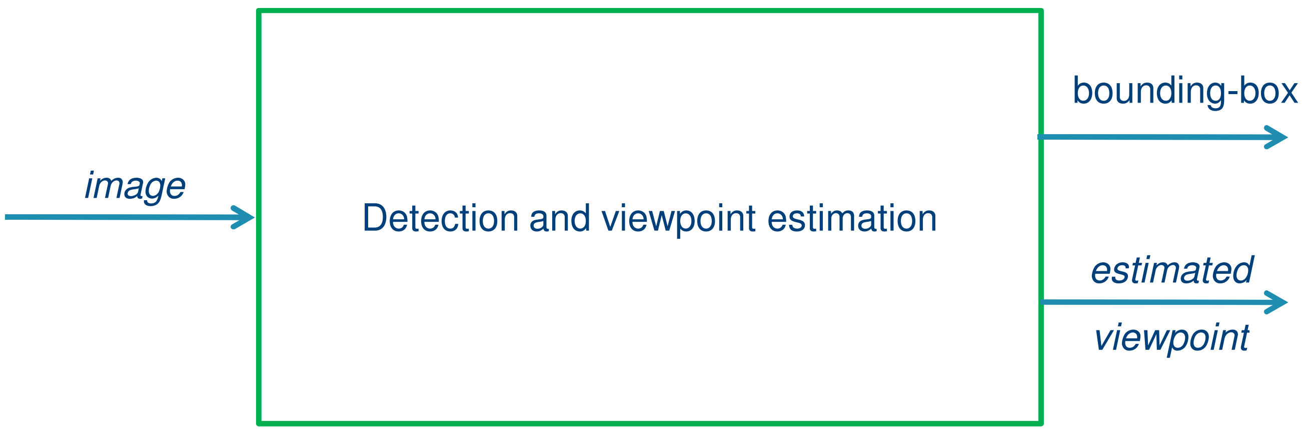
Related works: Chronological Orders

- Detector-based 2D models
- Detector-based 3D models
- Classification-based 2D models (current work)

PSI VISICS

KU LEUVEN

Common Pipeline

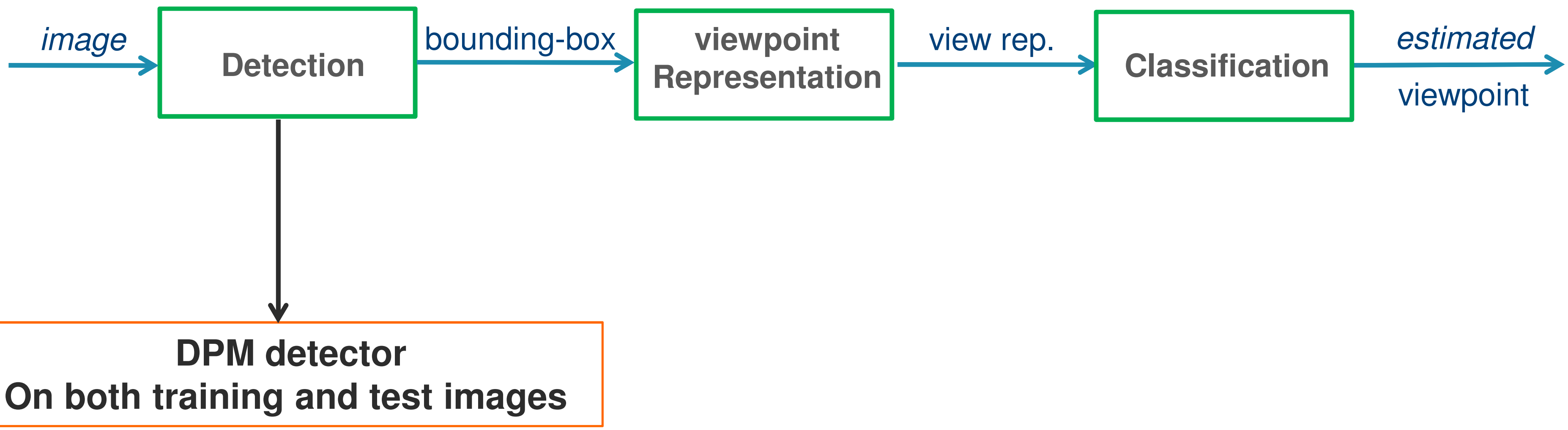


This pipeline is decomposed to

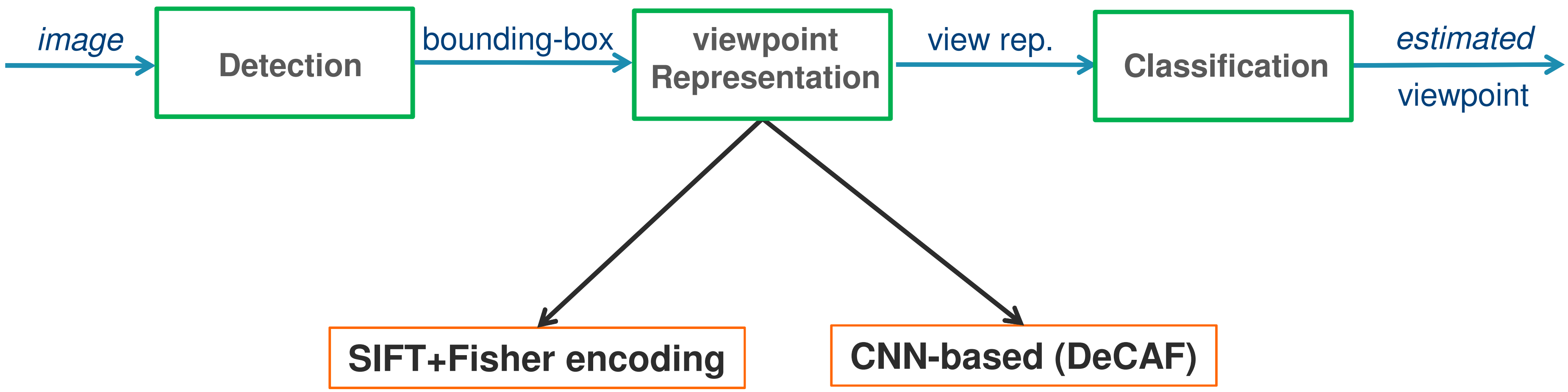
↓

Our Pipeline

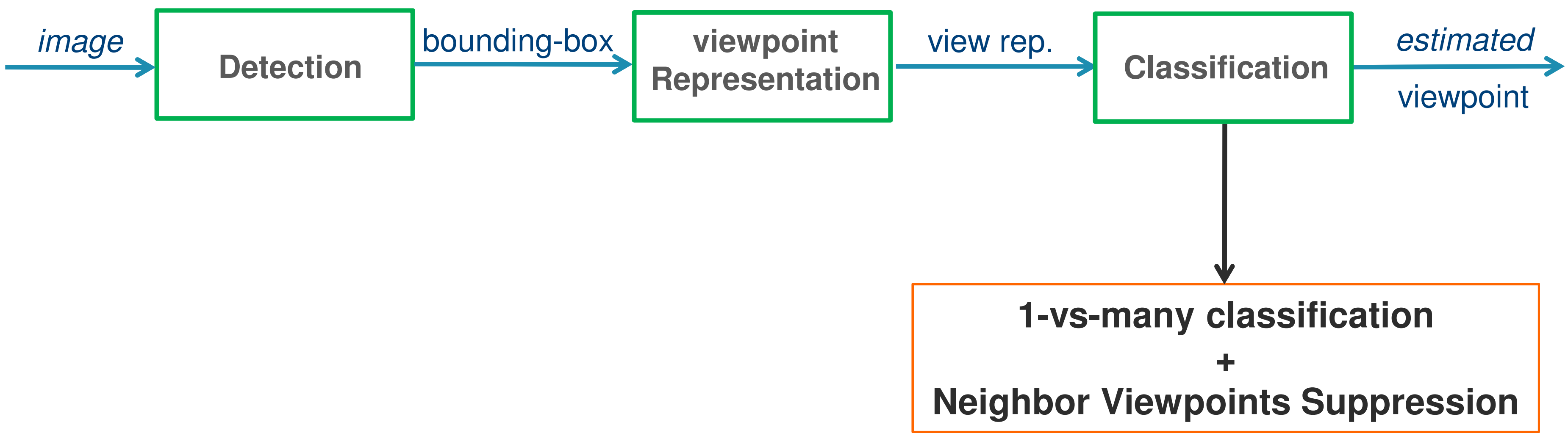
Our Pipeline



Our Pipeline

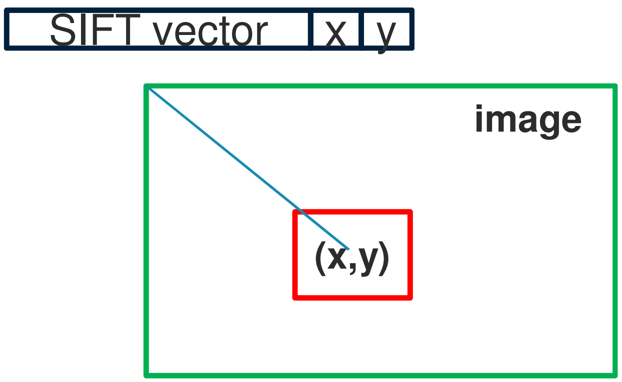


Our Pipeline



Enriching Fisher by Spatial Information

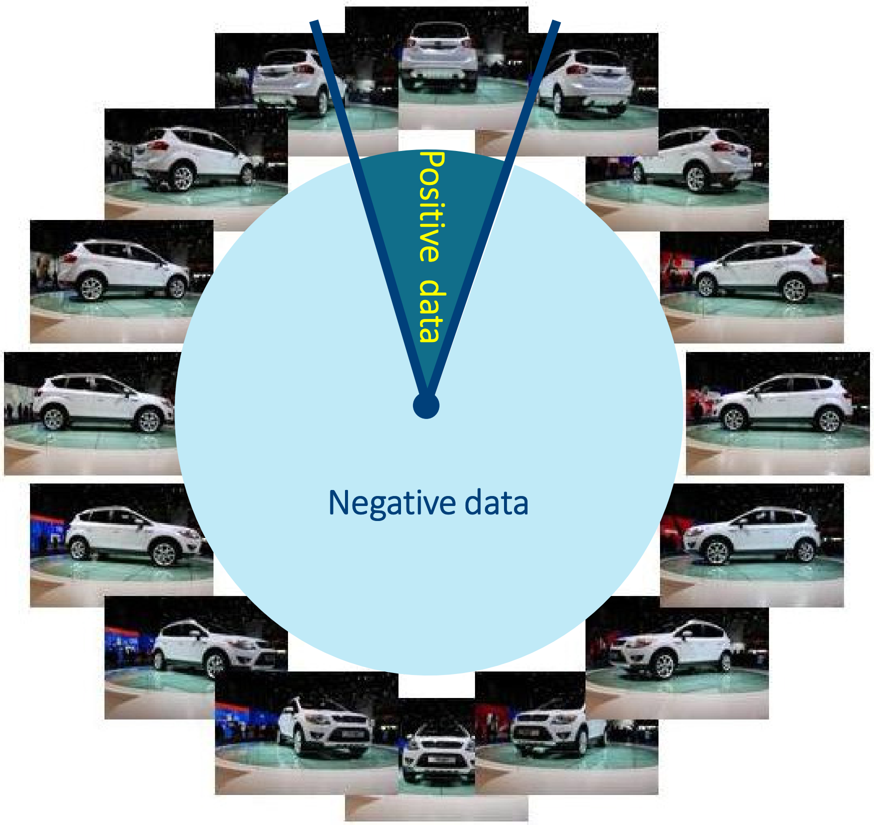
- Low-Level strategy
  - Augmenting dense SIFT with location of the patch.



- High-Level strategy
  - Building Spatial Pyramid of size 4×4, 2×2 and 1×1.

Learning

- Linear support vector machine classifier.
- Each viewpoint as a different class (1-vs-rest strategy).



Datasets - Cars

- Evaluated on EPFL multi-view car dataset
- 2299 images on **8/16/36** discretized viewpoints spanning over **360** degrees.

Characteristics: Fine binning of viewpoints, cars are in the center of images, no occlusion.

PSI VISICS

KU LEUVEN

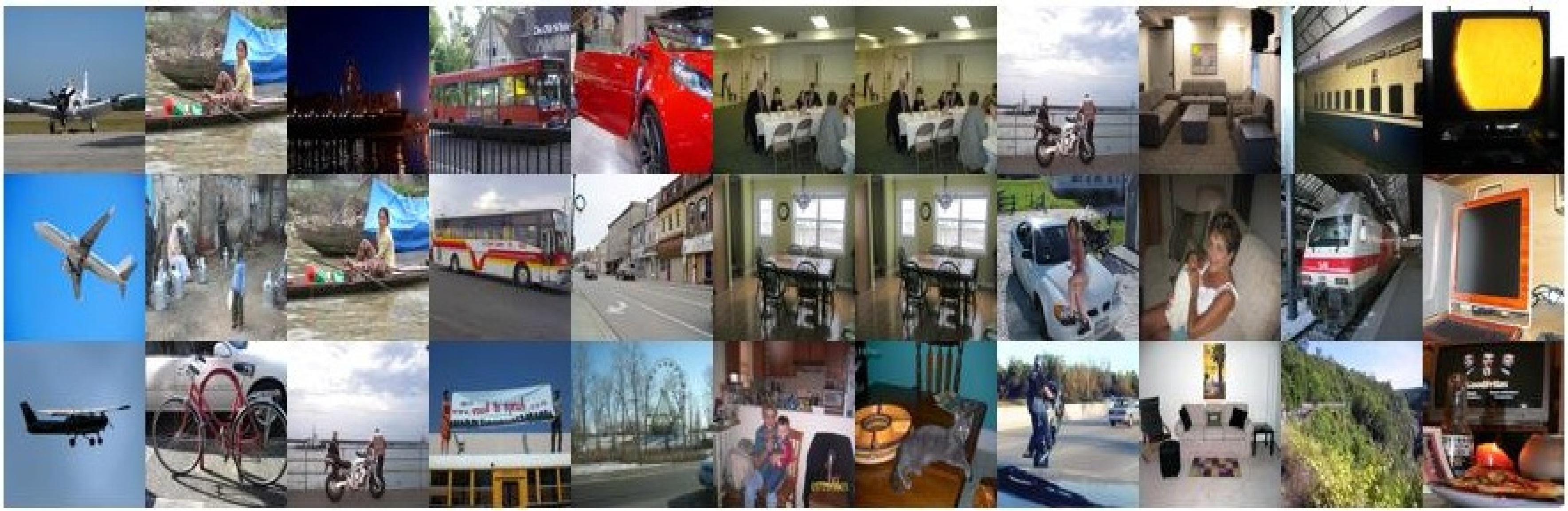
Datasets - Faces

- Evaluated on Annotated Faces-in-the-Wild (AFW) dataset.
- 468 faces, **13** discretized viewpoints spanning over **180** degrees.

Characteristics: Images contain cluttered backgrounds with large variations in face appearance

Datasets - General Objects

- Evaluated on PASCAL3D+ dataset.
- **11** rigid categories of PASCAL VOC 2012, **4/8/16/24** discretized viewpoints.



Characteristics: images exhibit much more variability.

PSI VISICS

KU LEUVEN

Results - Baseline

Bag-of-Words (BoW) representation is the poorest method.

		Cars (8 views)	Faces (13 views)
Feature Type	**Encoding**	**MPPE**	**FVP**
SIFT	BoW	54.8%	49.4%
SIFT	Fisher	68.2%	54.3%
SIFT	Fisher+SPM	80.1%	69.7%
SIFT+loc	**Fisher+SPM**	**81.8%**	**70.3%**
DeCAF	**-**	**72.0%**	**67.9%**

PSI VISICS

KU LEUVEN

Results - Baseline

Bag-of-Words (BoW) representation is the poorest method.

Best representation on both datasets is fisher with spatial pyramid (Fisher+SPM).

		Cars (8 views)	Faces (13 views)
Feature Type	**Encoding**	**MPPE**	**FVP**
SIFT	BoW	54.8%	49.4%
SIFT	Fisher	68.2%	54.3%
SIFT	Fisher+SPM	80.1%	69.7%
SIFT+loc	**Fisher+SPM**	**81.8%**	**70.3%**
DeCAF	**-**	**72.0%**	**67.9%**

PSI VISICS

KU LEUVEN

Results - Baseline

Bag-of-Words (BoW) representation is the poorest method.

Best representation on both datasets is fisher with spatial pyramid (Fisher+SPM).

Embedding spatial information in the low-level (SIFT+loc) is still advantageous.

		Cars (8 views)	Faces (13 views)
Feature Type	**Encoding**	**MPPE**	**FVP**
SIFT	BoW	54.8%	49.4%
SIFT	Fisher	68.2%	54.3%
SIFT	Fisher+SPM	80.1%	69.7%
SIFT+loc	**Fisher+SPM**	**81.8%**	**70.3%**
DeCAF	**-**	**72.0%**	**67.9%**

Results - Baseline

Bag-of-Words (BoW) representation is the poorest method.

Best representation on both datasets is fisher with spatial pyramid (Fisher+SPM).

Embedding spatial information in the low-level (SIFT+loc) is still advantageous.

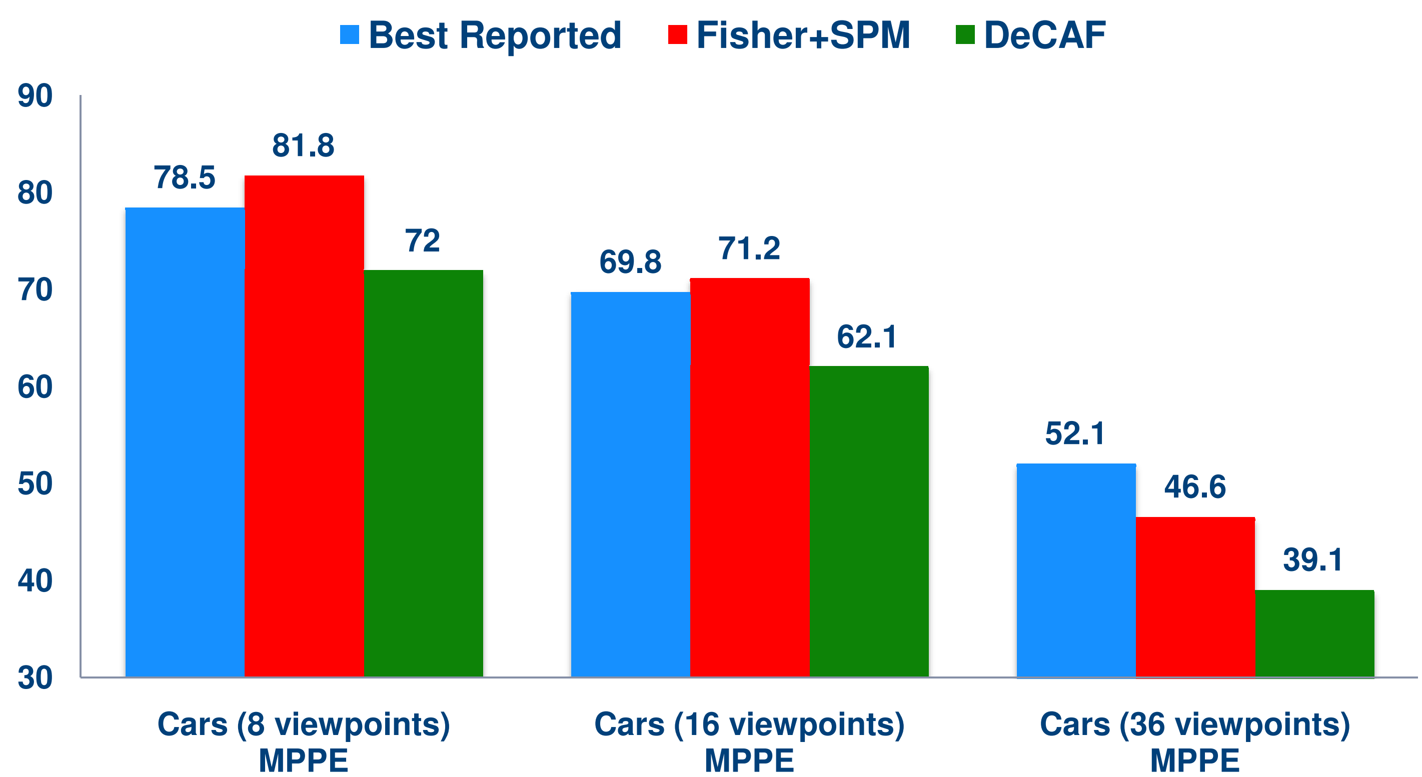
CNN-based features (DeCAF) performs quite good, especially considering their much lower dimensionality.

		Cars (8 views)	Faces (13 views)
Feature Type	**Encoding**	**MPPE**	**FVP**
SIFT	BoW	54.8%	49.4%
SIFT	Fisher	68.2%	54.3%
SIFT	Fisher+SPM	80.1%	69.7%
SIFT+loc	**Fisher+SPM**	**81.8%**	**70.3%**
DeCAF	**-**	**72.0%**	**67.9%**

PSI VISICS

KU LEUVEN

Cars - Comparison with state-of-the-art

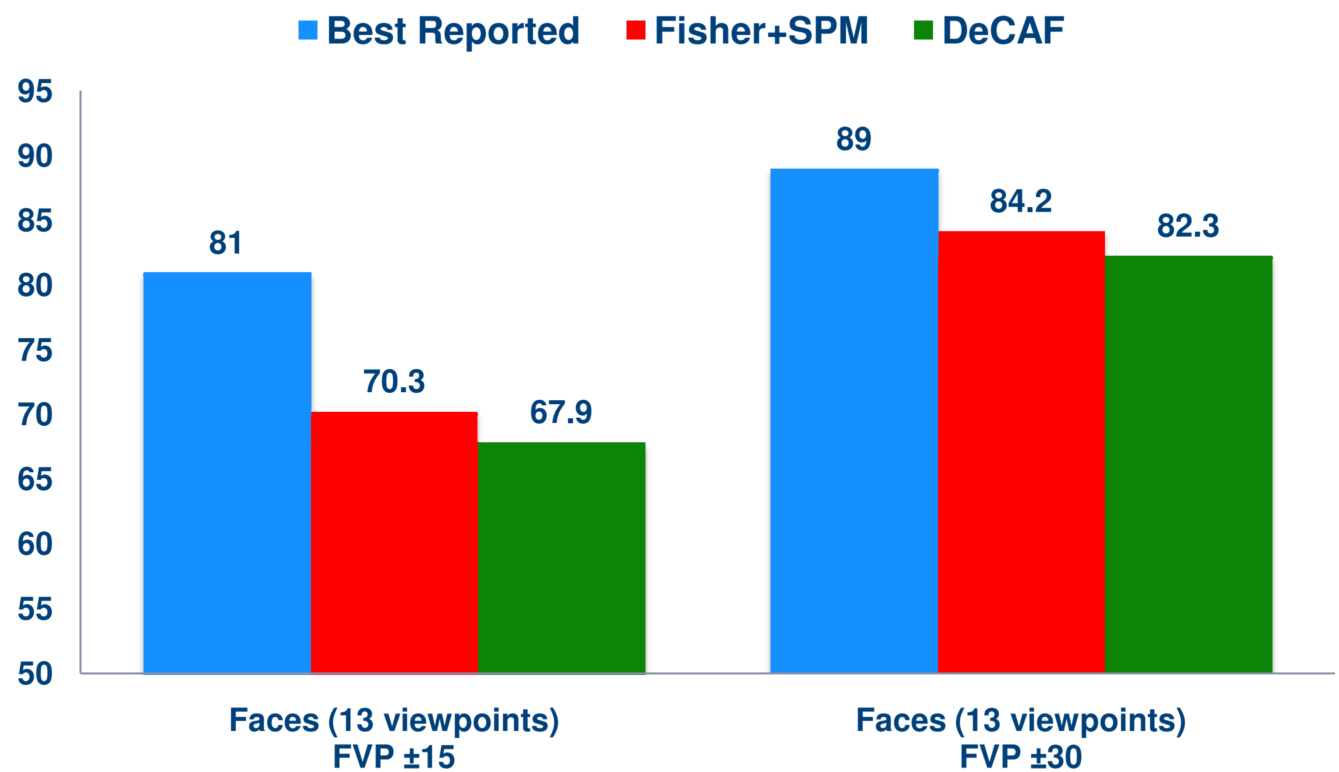


- (Left) B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3d2pm–3d deformable part models. In ECCV, 2012

PSI VISICS

KU LEUVEN

Faces - Comparison with state-of-the-art



X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In CVPR, 2012

PSI VISICS

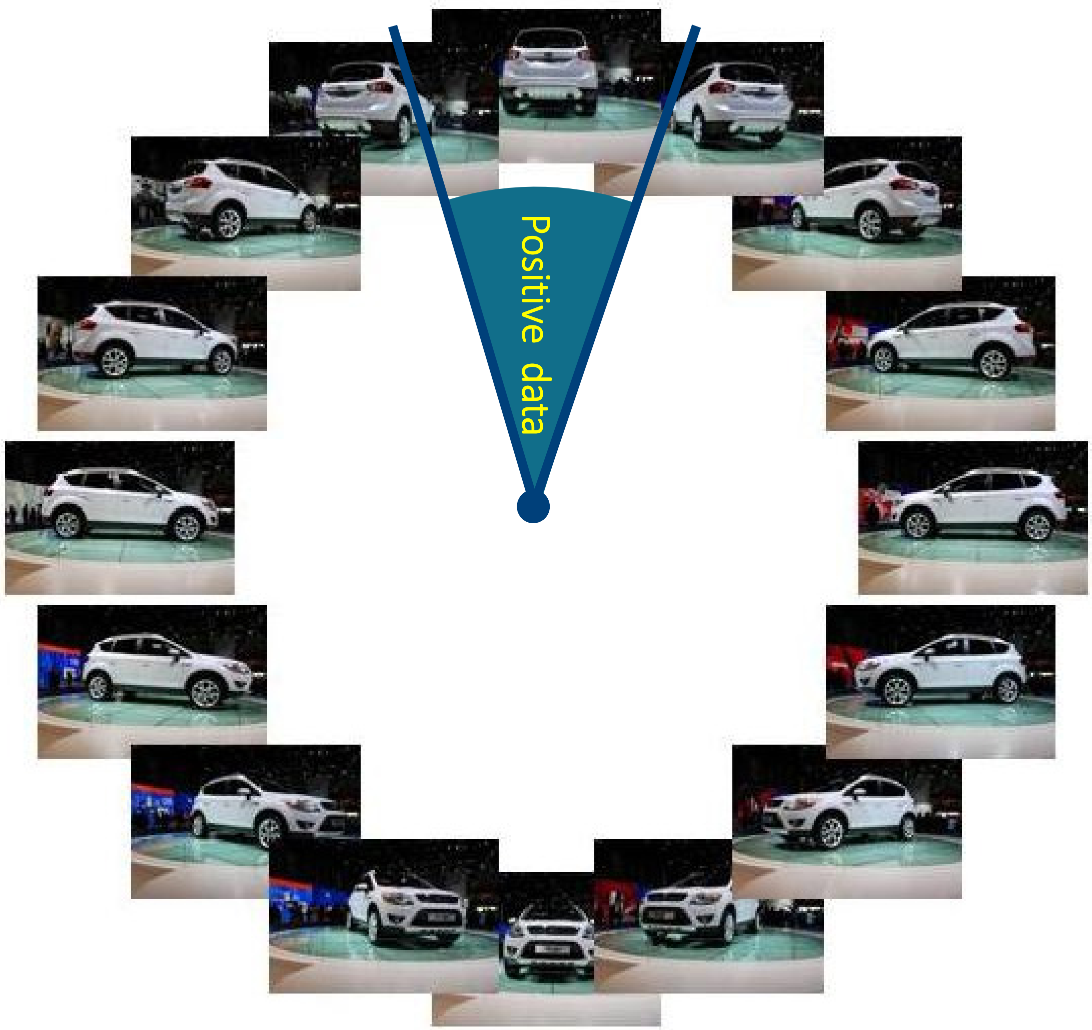
KU LEUVEN

Learning - Challenges

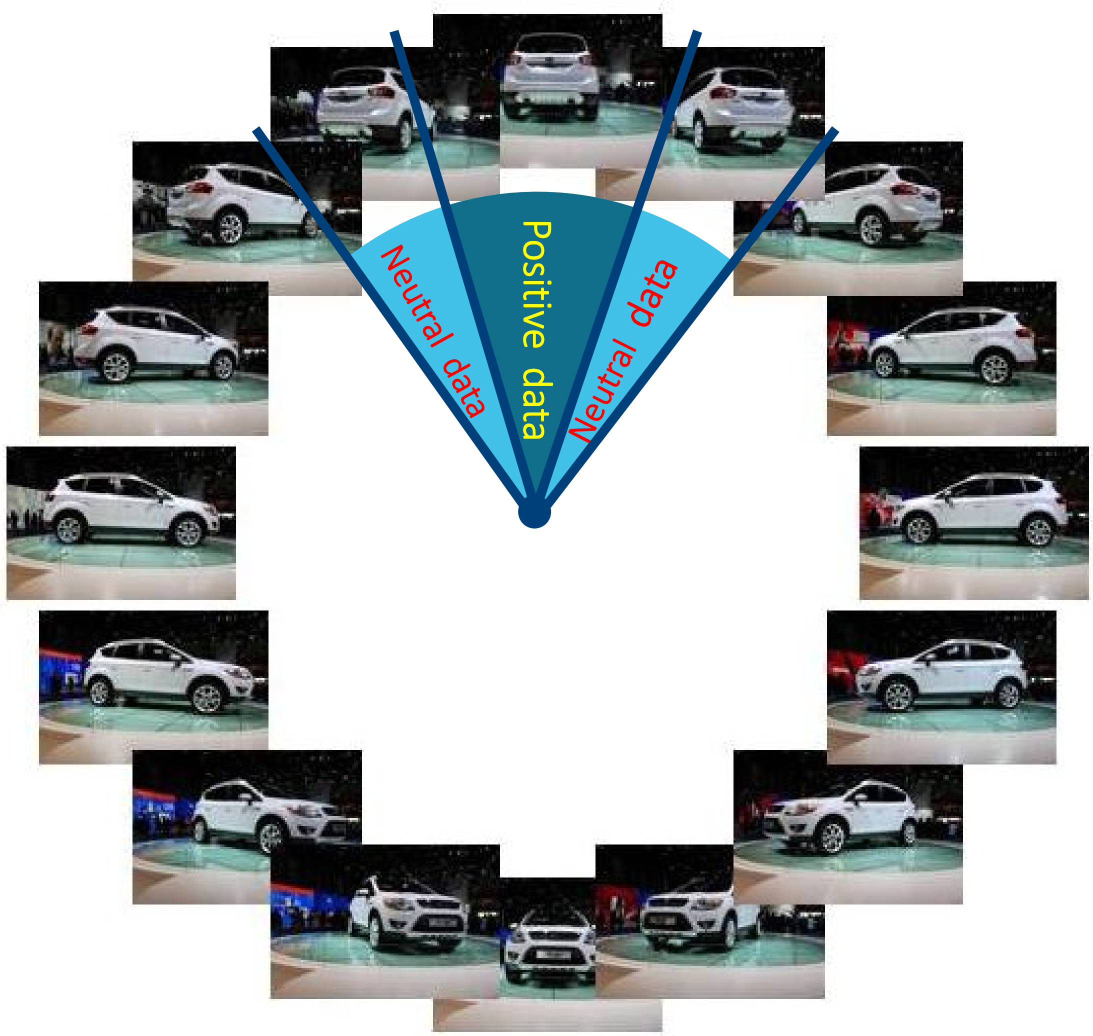
- Nearby viewpoints are visually very correlated.

- Classifier mainly focuses on distinguishing positive viewpoint from similar nearby viewpoints.

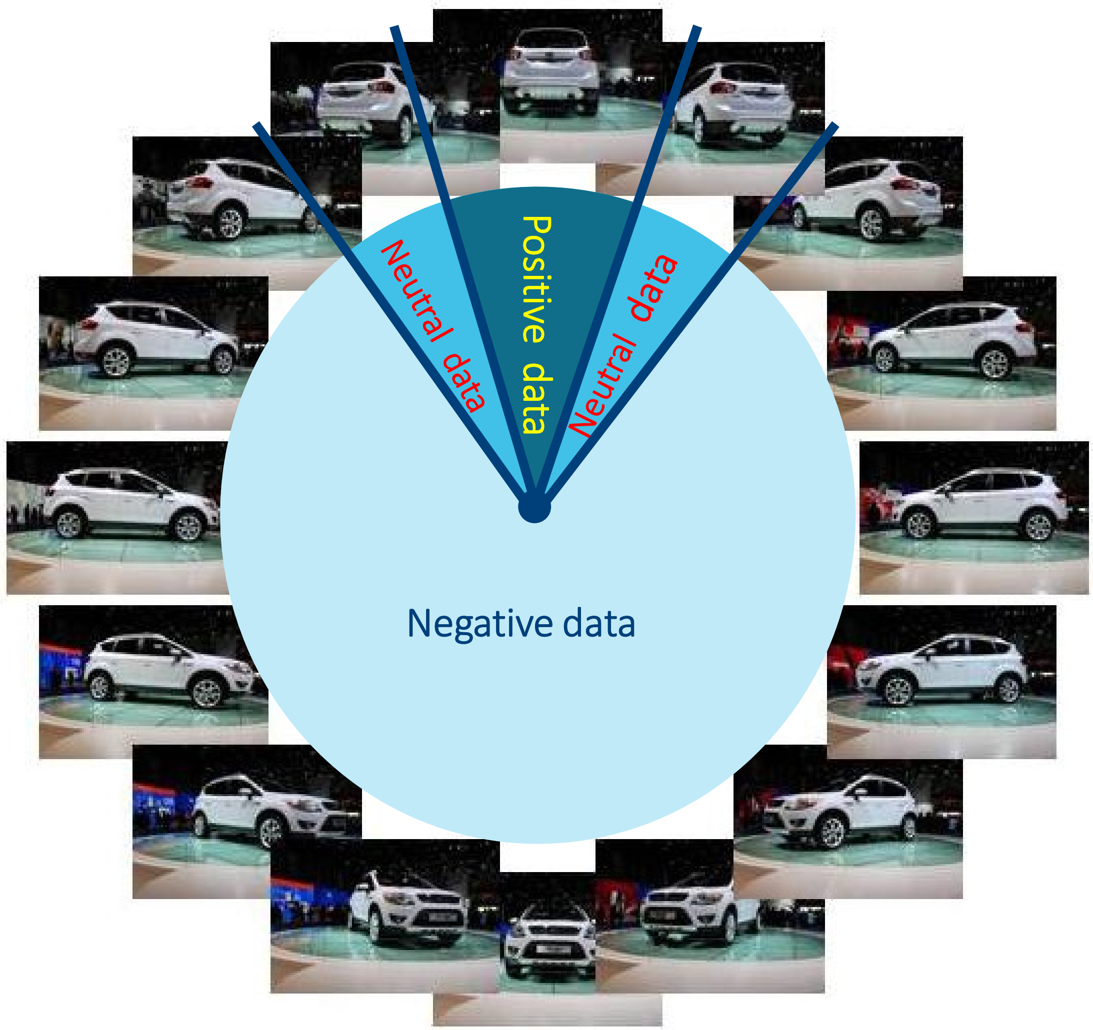
Neighbor Viewpoints Suppression



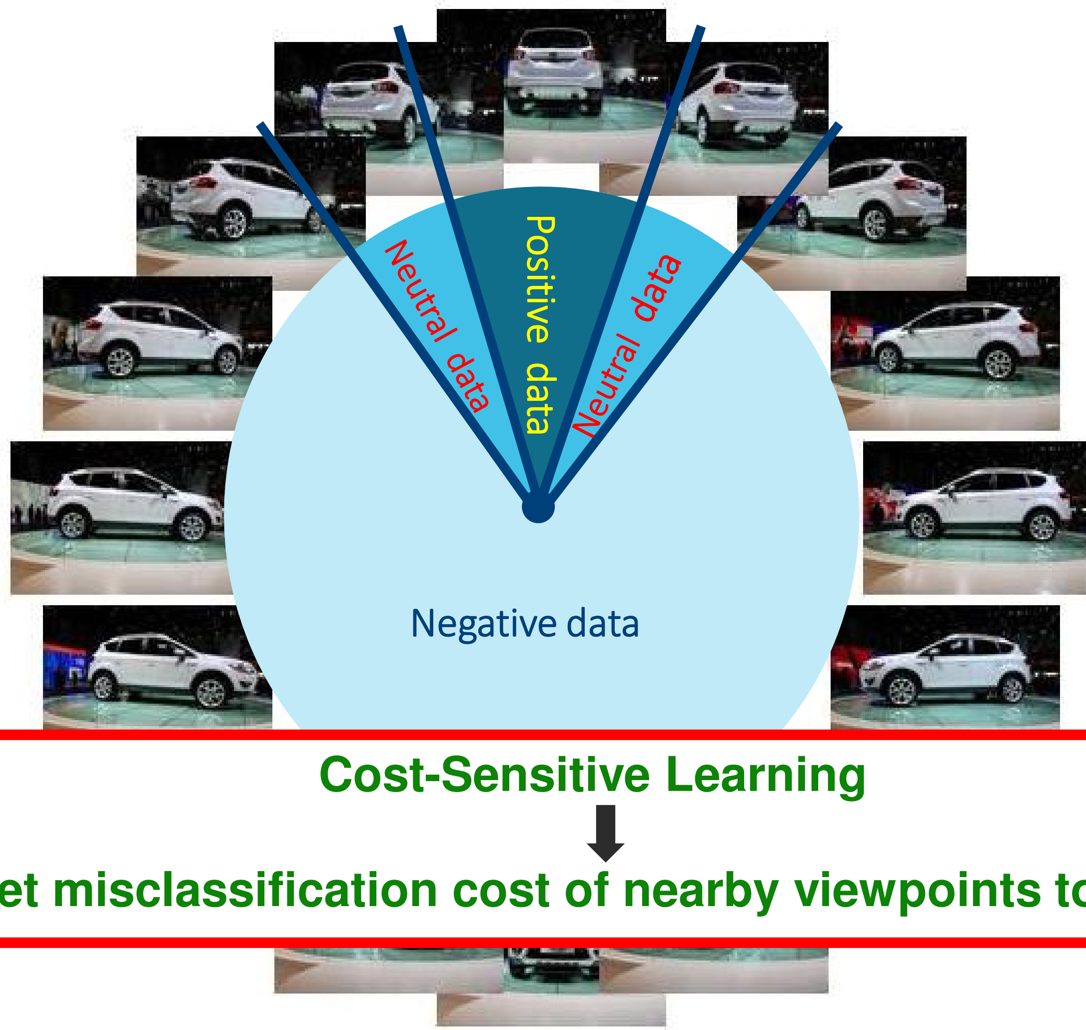
Neighbor Viewpoints Suppression



Neighbor Viewpoints Suppression



Neighbor Viewpoints Suppression



Cost-Sensitive Learning

⬇

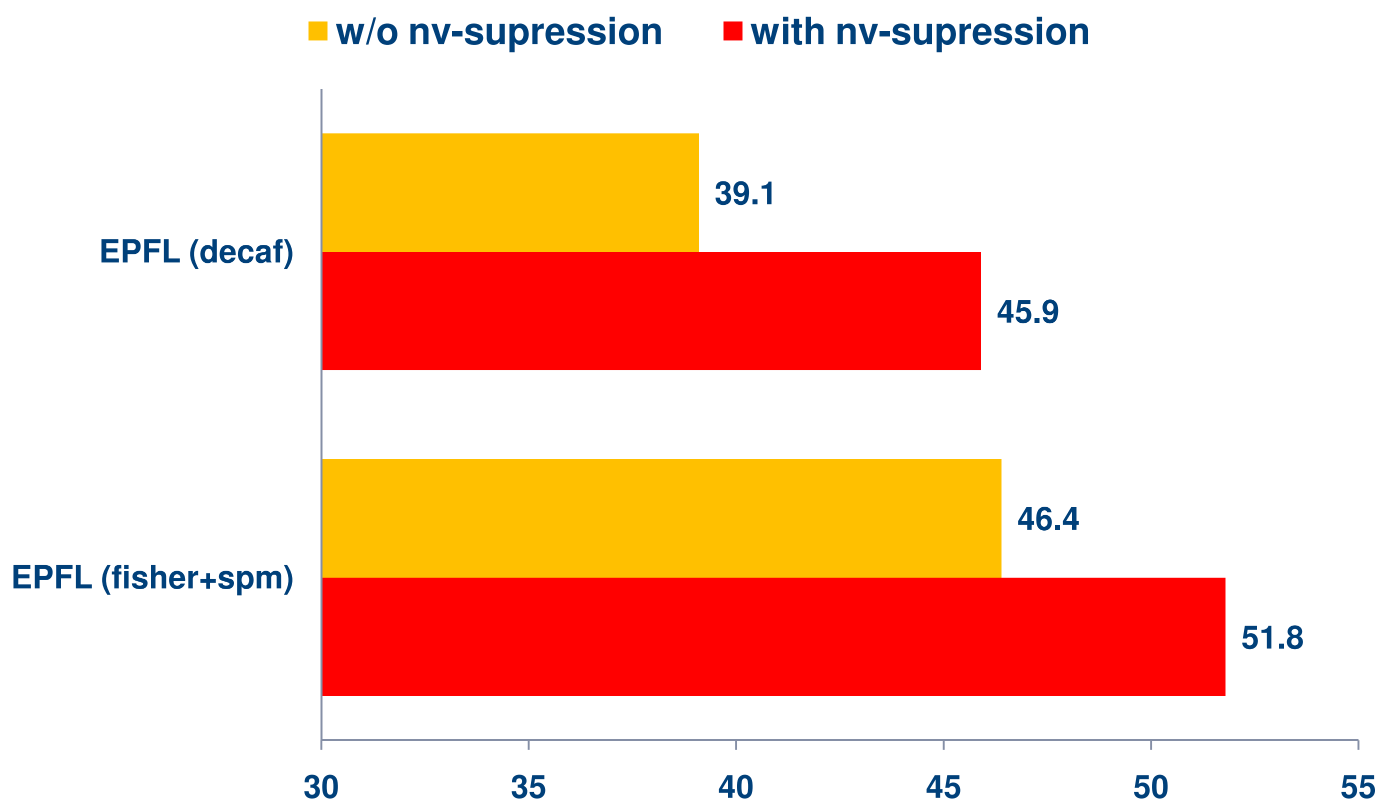
Set misclassification cost of nearby viewpoints to zero

PSI VISICS

KU LEUVEN

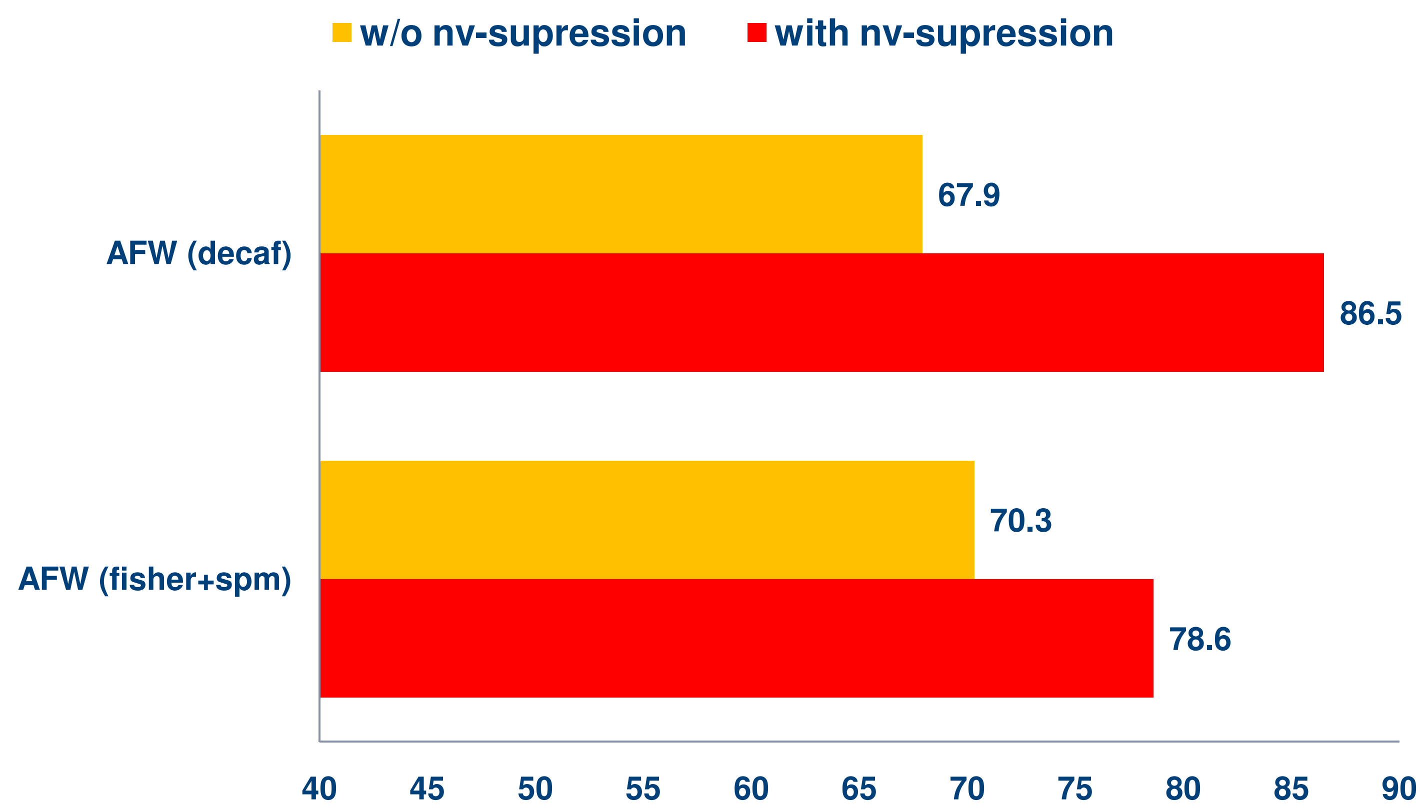
Results – Neighbor Viewpoints Suppression

EPFL cars dataset – 36 bins

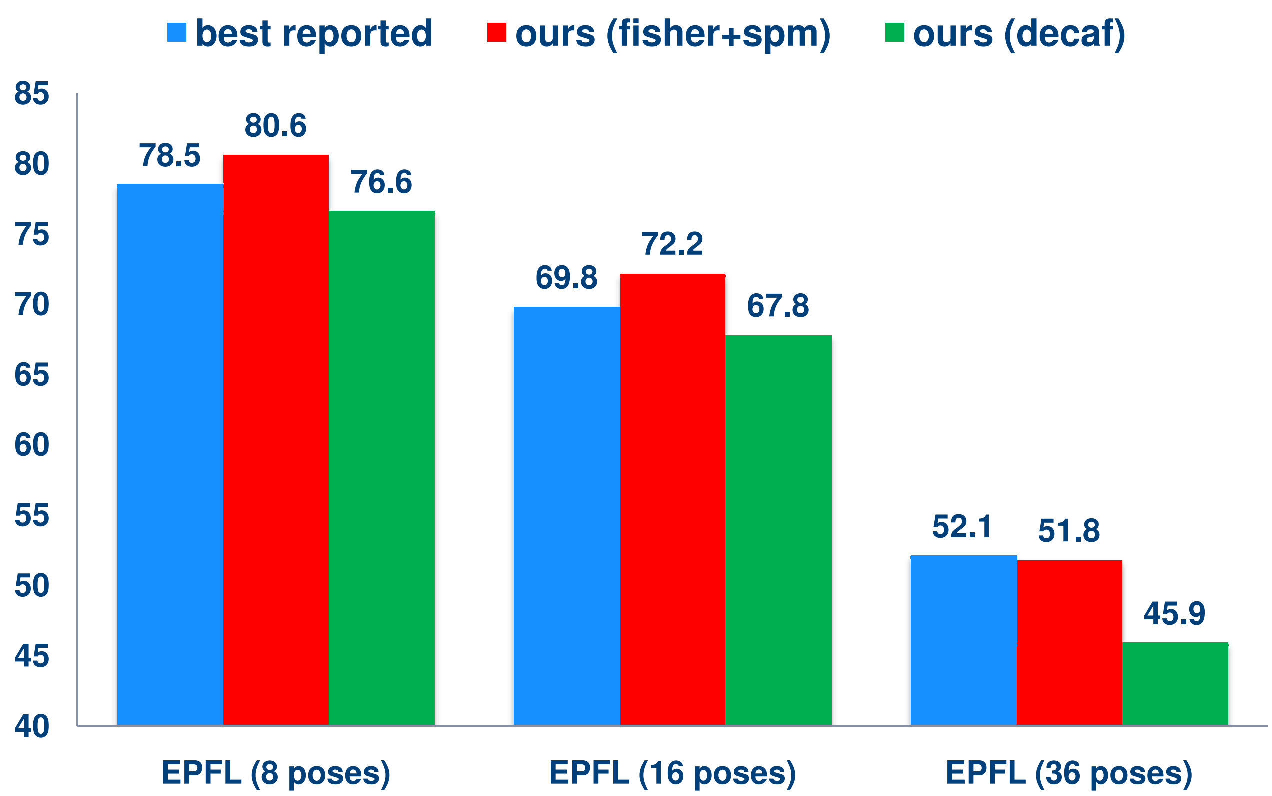


Results – Neighbor Viewpoints Suppression

AFW faces dataset – 13 bins



Cars - comparison with state-of-the-art

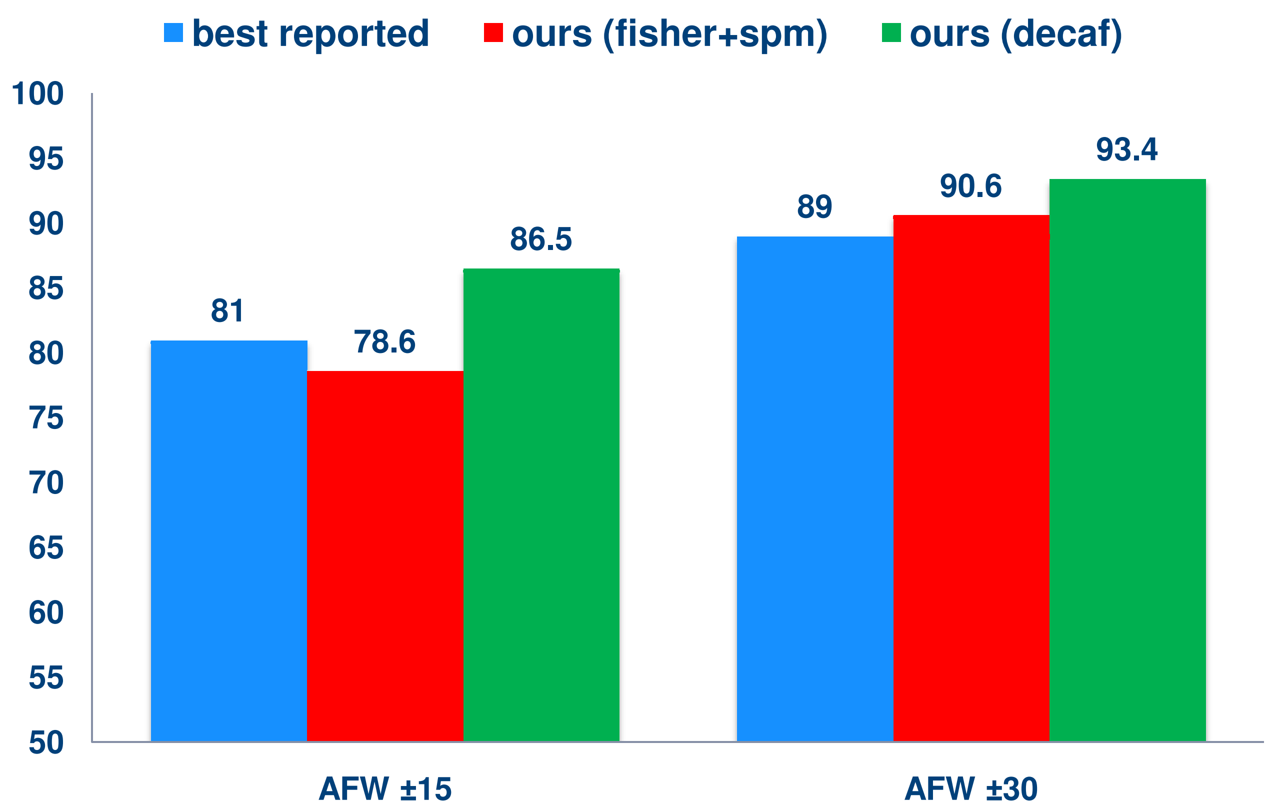


- B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3d2pm–3d deformable part models. In ECCV, 2012

PSI VISICS

KU LEUVEN

Faces - comparison with state-of-the-art

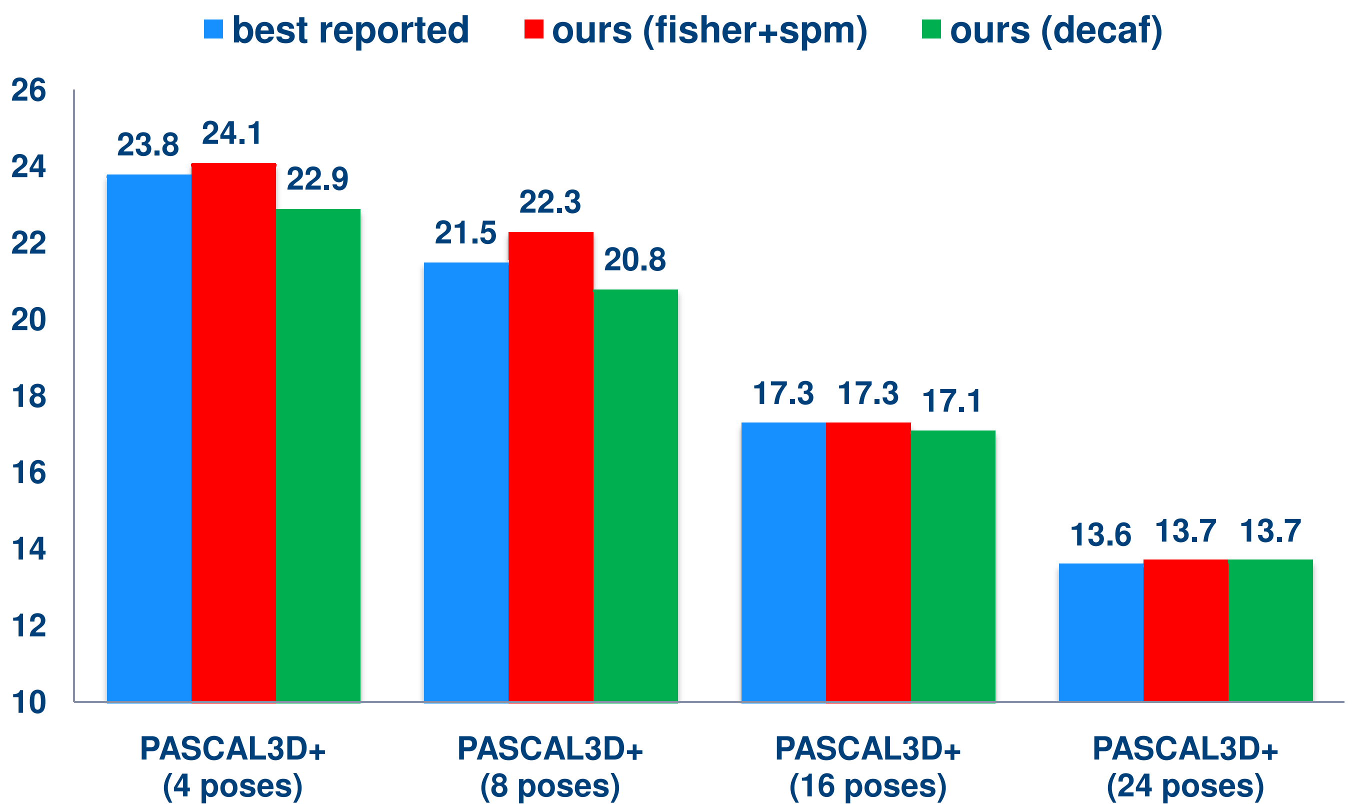


- X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In CVPR, 2012

PSI VISICS

KU LEUVEN

Objects - comparison with state-of-the-art



- B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In CVPR, 2012.

PSI VISICS

KU LEUVEN

Computational Costs

- Time complexity of our pipeline

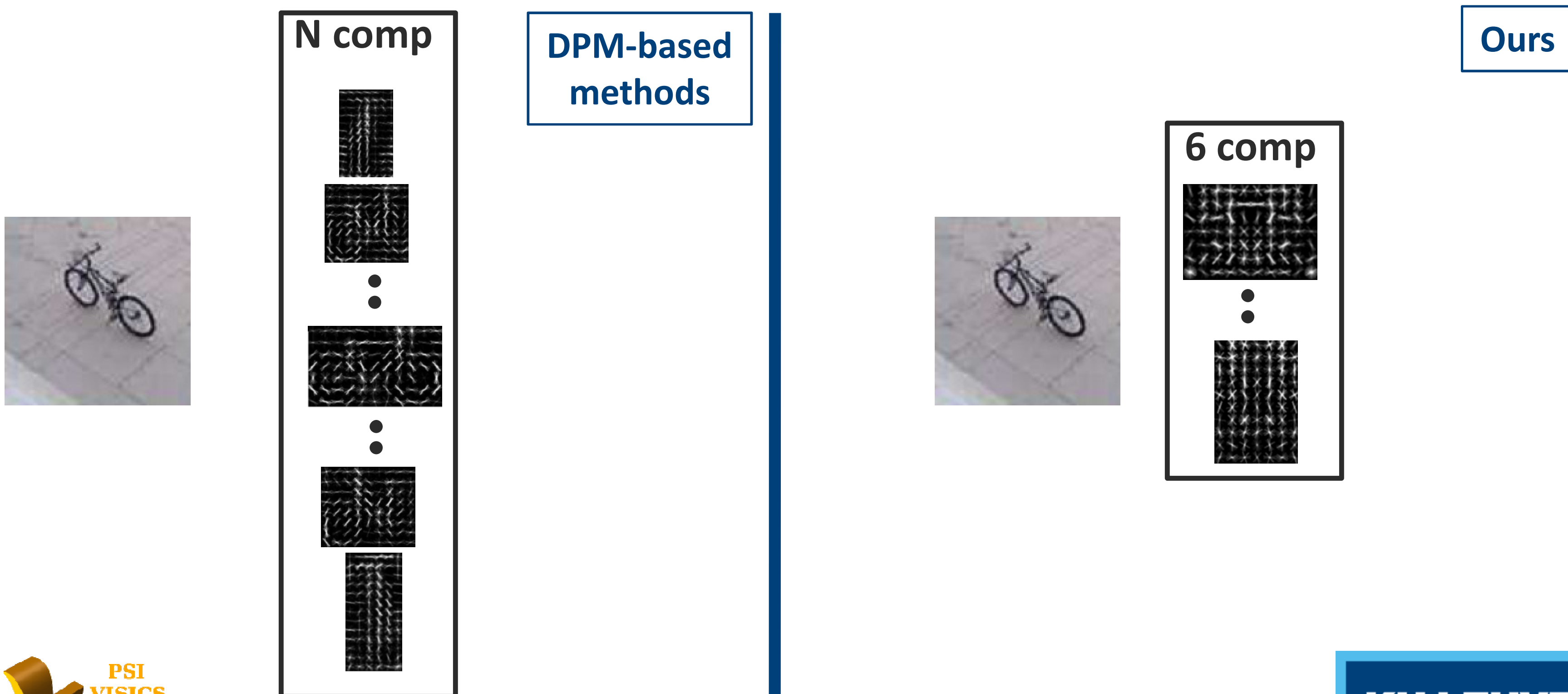
EPFL dataset	
Task (per image)	Average time (sec)
Detection	4
Extracting SIFT + Fisher vector pyramid	2
DeCAF feature extraction	0.2
36-bins view classification	0.19

PSI VISICS

KU LEUVEN

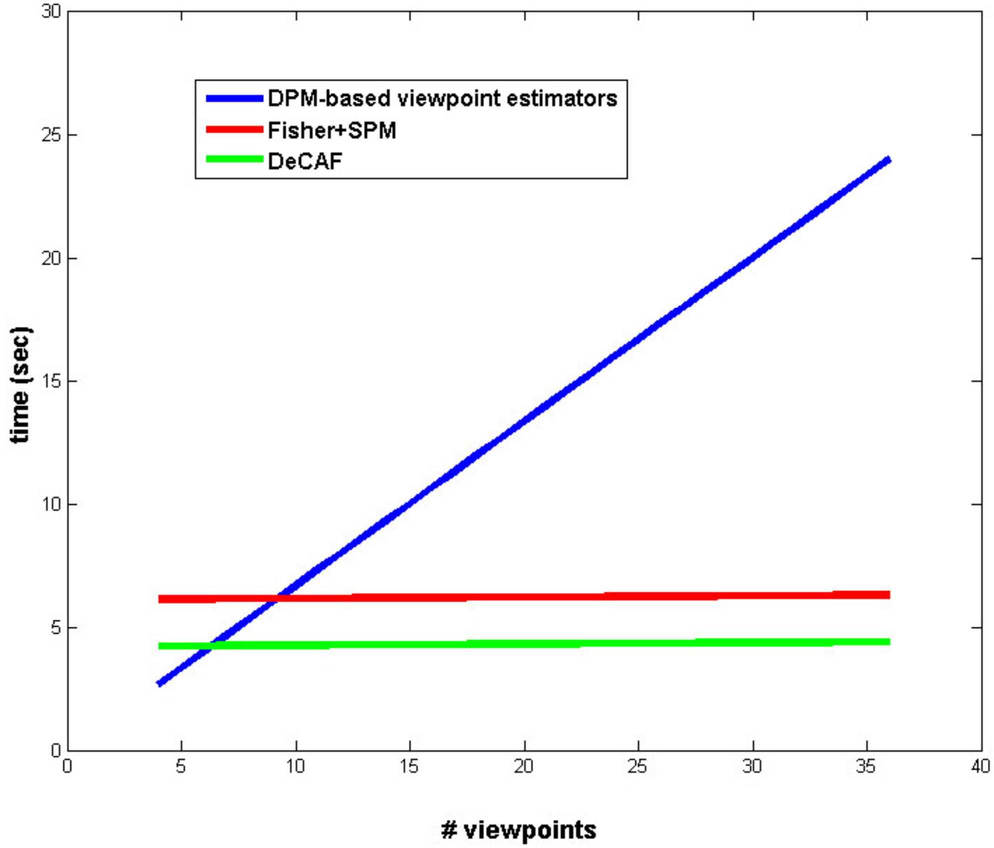
Computational Costs

- We can safely claim that all the methods based on DPM are computationally more demanding.
    - we use standard DPM models with 6 components while others generally use a DPM component for each view.



Computational Costs

- We can safely claim that all the methods based on DPM are computationally more demanding.
  - we us
    enerally use a
    DPM



Ours

Conclusion

- We have presented a study of different methods for view estimation.
- In contrast to common believe, the very simple 2D framework, if properly tuned, can in most of the cases outperform the state-of-the-art including methods based on 3D or more complex and computationally expensive models.
- It suggests the next generation of view estimation methods should probably combine these powerful 2D representations with 3D reasoning.

Thanks For Your Attention!
Questions?

Outline

- Problem Definition
- Related works
- Pipeline
- Datasets and Evaluations
- Conclusion

Discussion

- Considering that DeCAF and Fisher are general representations and are not designed specifically for the viewpoint estimation problem, they surprisingly performs well.

- On EPFL cars and PASCAL3D+ dataset, Fisher performs better than DeCAF, while in AFW faces, DeCAF surprisingly performs better after applying neighbor viewpoint suppression procedure.

- The advantage of DeCAF is its lower dimensionality compared to Fisher+SPM.

Computational Costs

- Time complexity of our pipeline

EPFL dataset	
Task (per image)	Average time (sec)
Detection	4
Extracting SIFT + Fisher vector pyramid	2
DeCAF feature extraction	0.2
36-bins view classification	0.19
Training 36 one-vs-rest linear SVM	290

Standard 1-vs-rest Classifier